AZELLA Kindergarten Placement Test Score Validity Documentation

June 2015

Lietta Scott, Jacqueline Church, and Jessica Eilertson

Arizona Department of Education – Assessment Unit

Executive Summary

This report is to provide stakeholders and policy makers with an overview of the evidential argument that the Arizona Department of Education (ADE) has developed in support of the valid use of the AZELLA Kindergarten Placement Test (Placement Test) scores. The test was designed and developed for use with incoming Kindergarten students who have a primary home language other than English (PHLOTE), and provides the criterion for eligibility to receive English Language Learner services.

The report is broken down into two sections plus an appendix. The first section gives a brief description of what is meant by test score validity. Test score validity pertains to the interpretation and use of a test's scores. A test is not valid or invalid; rather, the interpretation and use of the test scores (results) are valid or invalid for a specific purpose. In order to make the claim that a test's scores are valid, the test developers must provide evidence. This is done through a validation process, which consists of developing an argument that provides evidence to support the test's use and interpretation. This argument consists of many pieces, and is an ongoing process, and as such, is a matter of degree rather than an absolute.

The second section summarizes the evidence that ADE has gathered for the validity argument of the interpretation and use of the Placement Test scores as appropriate for placing PHLOTE students into an instructional setting. The evidential components described in this summary include: administration and scoring, documentation, educational tests/ policy and accountability, fairness, reliability and accuracy, scores and scales, test design and development, users' responsibility, and construct validity.

The appendix provides a detailed list of the evidence by the evidential component to which it aligns. These were developed by ADE using well-established practices in the field of assessment and are structured based on work of Mislevy, Steinberg, and Almond (2003), Mislevy (2004), Bachman & Palmer (2010), and Chapelle, Enright, & Jamieson (2008). The argument structure frames the interpretations, conclusions, or claims that are drawn about the test takers' abilities. A link to the document on ADE's website is listed for each piece of evidence in this third section. A glossary of the terms used in the evidence structure is also provided.

AZELLA Kindergarten Placement Test: Validity

Tests seem to be universal in our society; there are college entrance exams, state standardized tests, tests of athletic ability, such as the scoring of a diving competition, and so on. A test is created to measure a construct, or specific elements of a person's knowledge, skills, and/or abilities. For example, a Grade 5 reading test is designed to determine a 5th grade student's reading ability and the student's score might be used to aid in deciding what support the student needs. This example shows the most important aspect of testing: the way in which test results are interpreted and used.

This notion of a test's interpretation and use is what is meant by validity. A test cannot be valid or invalid, only the interpretation and use of its scores (results) can be valid or invalid for a specific purpose. Test score validity means that the interpretation and use of a test is valid, or appropriate, for a specific purpose. For example, a test that is meant to measure a student's ability in reading would not appropriately be used to recommend students for a school's mathematics competition team.

Validity is generally thought of as the most important consideration in educational testing because it has to do with how test scores are used. When a person says that a test is valid, what they really mean is that the interpretations of the test results for a specific purpose are valid. An example of this would be a test to join a track and field team that requires a runner to run a mile in less than five minutes. This running test matches the intended use of the test results. This example illustrates validity for a simplified test; however, most tests used in education are much more complex and require a more rigorous process to demonstrate validity.

In order to make the claim that a test's scores are valid, the test developers must provide evidence. This is done through a validation process, which consists of developing an argument that provides evidence to support the test's use and interpretation. This argument consists of many pieces, one of which generally includes examining the fairness of the test. For this part of the argument, any evidence to explain the fairness of the test is collected and used as evidence. Some of the information, like considering potential issues of students' rights, is considered before the test is administered, while other information, like reliability, is analyzed after the test is administered. Thus, test score validation is an ongoing process, and as such, is a matter of degree rather than an absolute.

The interpretation and use of the AZELLA Kindergarten Placement Test (Placement Test) scores is only appropriate for placing students, who have a primary home language other than English, into an instructional setting. Scores on the Placement Test determine whether students meet the criteria for English Language Learner (ELL) services. The components used in developing the evidence for the Placement Test validity argument include: administration and scoring, documentation, educational tests/ policy and accountability, fairness, reliability and accuracy, scores and scales, test design and development, users' responsibility, and construct validity. These components will be examined in more detail in the following pages.

AZELLA Kindergarten Placement Test: Validity Evidence

Following are evidential components that contribute to the Placement Test score validation process. The descriptions of these elements provide an explanation of the considerations that were made by the Arizona Department of Education (ADE) during the test design, development, scoring, and scaling process. Together they contribute to the evidential argument of the validity of the Placement Test's scores for informing decisions of educational placement for incoming Kindergarten students who have a primary home language other than English.

Administration & Scoring: Test administrators are provided with test administration documents as well as training to facilitate the administrations of all tests in a manner that keeps construct-irrelevant influences to a minimum. Test administration guides are standard across the state to ensure that all tests are administered in the best possible and most consistent manner. These guides are reviewed prior to each new school year and updated as necessary. Rubrics used for scoring reflect the English Language Proficiency Standards (ELPS). In order to administer and score the Placement Test, scorers must take an approved training and pass a qualifying quiz annually. A purpose of this training is to ensure reliability.

Documentation: Multiple technical and research reports are available publically on the ADE Assessment section website at http://www.azed.gov/assessment/technical-legal-resources/.

Educational Tests/ Policy & Accountability: ADE has contracted with the National Center for Educational Outcomes to conduct an investigation into the consequences of the test for students who score Proficient on the Placement Test. This study is being conducted during the 2014-2015 school year. Results will be forthcoming and will inform the validity of the Placement Test's scores.

Fairness: Fairness is an important consideration in testing. Fairness includes many issues, such as ensuring equitable treatment of all examinees. Bias and sensitivity meetings were held for the Placement Test to ensure that the assessment is fair and accessible to students. For the Placement Test, test administrators are provided with test administration documents and training to make certain that all tests are administered in a way that eliminates any construct-irrelevant characteristics. The Placement Test should have a positive impact on English teaching in both ELL and mainstream classrooms. A longitudinal study of students assessed with the placement test is underway. This study's goal is to affirm the claim that the test makes valid score interpretations for examinees from specific subgroups. Once students are identified, there are a variety of options that can be used, such as different models for instruction. The test establishes the criterion for eligibility to receive services.

Reliability and Accuracy: Reliability is a measure of the degree to which a test measures what it purports to measure. Reliability on the Placement Test is computed using a measure of internal consistency (Coefficient Alpha). The reliability statistic for the Placement Test is 0.95 out of 1.00. Additionally, the decision consistency statistic and the accuracy statistic is 0.91 and 0.94 out of 1.00, respectively. These metrics indicate that the Placement Test does a good job in providing information about the language ability of most incoming Kindergarten students.

Scores & Scales: Decisions about cut scores were made by a standard setting committee comprised of educators with expertise with both mainstream and ELL Kindergarten students. Scores are tied to the Performance Level Descriptors (PLDs), and score interpretation information is provided on the

student report. Student reports show the students' overall score, performance level, and a brief description of each performance level. The *Guide to Navigating and Using AZELLA Reports* (http://www.azed.gov/assessment/files/2014/08/report-use-8-21-14.pdf) gives instructions for understanding and using the reports. Placement test scales are monitored to ensure that the scoring scale is stable.

Test Design & Development: The Placement Test was designed and developed to measure social and academic language. ADE operationalized this construct using the following definition:

> [Academic language is] the language students need to meaningfully engage with academic content within the academic context…academic language includes the words, grammatical structures, and discourse markers that facilitate student access to and engagement with grade-level academic content.[1]

The test was also informed by the ELPS and Early Learning Standards. The language being assessed on the test has been identified as necessary for incoming Kindergarten students in the state of Arizona. The item specifications and test blueprints support the measurement of language skills and knowledge described in the construct for all language proficiency levels and language modalities. Test and item specifications and blueprints were created by test experts and educators who have expertise with incoming Kindergarten students who have a primary home language other than English. During the development of the Placement Test, a cognitive study was carried out to evaluate proposed tasks and process. Item statistics were reviewed to verify appropriateness of use. The tasks that worked well were included on the test. The test construction process, test administration instructions, process for scoring of items, item analysis results and calibration, equation, and scoring information can be found in the *Technical Report* on ADE's website at: http://www.azed.gov/assessment/files/2014/06/azella_tech_report_2012-2013_final.pdf.

Users' Responsibilities: The test results are intended to be used to make decisions about whether incoming Kindergarten students are eligible for ELL services. The results also inform schools' and teachers' decisions about the appropriate curriculum for students. Test use is delegated to those who have the training, credentials, and experience required to appropriately interpret and use the data. Because of the high-stakes nature of score interpretations for students, test security is of the utmost importance and is carefully monitored.

Construct Validity: The Placement Test assesses the construct of social and academic language through the use of the Arizona ELPS and tests the domains of listening, reading, writing and speaking. The skills and abilities assessed on the test are reflective of the ELPS and PLDs. The proficiency levels – Pre-Emergent/Emergent, Basic/Intermediate, and Proficient – reflect the level the student has attained on the ELPS. A score of proficient indicates that the student has a sufficient working knowledge of English to be able to access mainstream classroom instruction in English. Those students scoring less than proficient are identified as ELL and are to be provided with targeted instruction to support English language development.

[1]: Framework for high-quality English language proficiency standards and assessments. Prepared by the Assessment and Accountability Comprehensive Center, January 2009.

References

Bachman, L., & Palmer, A. (2010). *Language assessment in practice*. Oxford: Oxford University Press.

Chapelle, C. A., Enright, M. K., & Jamieson, J. M. (2008). *Building a validity argument for the test of English as a foreign language.* New York, NY: Routledge.

Mislevy, R. J. (2004). Toulmin and beyond: Commentary on Michael Kane's "Certification testing as an illustration of argument-based validation," *Measurement: Interdisciplinary Research and Perspectives*, *2*(3), 185-191.

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2003). On the structure of educational assessments. *Measurement: Interdisciplinary Research and Perspectives*, 1, 3-62.

Appendix A

AZELLA Kindergarten Placement Test: Current Evidence for Validity Argument

In developing the validity argument for the Placement Test, ADE used well-established practices in the field of assessment. The argument structure for the Placement Test was based on the work of Mislevy, Steinberg, and Almond (2003), Mislevy (2004), Bachman & Palmer (2010), and Chapelle, Enright, & Jamieson (2008). The argument structure frames the interpretations, conclusions, or claims that are drawn about the test takers' abilities.

The claims being made for the test's interpretation and use are described in detail in the following table. The rows in the table represent each claim, and the columns provide evidence for that claim. The table, as presented, is a working document and as such, is subject to revision as more evidence is gathered. A glossary of the terms used within the table is provided below.

Claim: conclusions drawn about the test; require justification

Warrant: a generally held principle or established procedure that authorizes the inference

Inference: logical conclusions that can be drawn based on the evidence

Assumptions Underlying Warrant: to examine the inference, a number of assumptions must be made, e.g., the assessment tasks are representative of the academic domain

Examples of Backing: a body of knowledge or evidence, scientific theories, or precedents as relate to the warrant and claim

Backing for Assumptions: evidence that justifies the assumptions that underlie the inference

Links to Documentation: links to the technical reports that provide additional information for each of the claims

Current Evidence for Validity Argument of AZELLA Kindergarten Placement Test
March 23, 2015

Claim	Warrant	Inference	Assumptions Underlying Warrant	Examples of Backing	Backing for Assumptions	Links to Documentation	Page Number(s)
Observation	Observations of performance on test reveal the relevant knowledge, skills, and abilities in situations representative of those in the target domain of language use for incoming Arizona Kindergarteners.	Domain Description	Critical English language skills, knowledge and processes needed for study in Kindergarten can be identified.	Domain Analysis	Blueprint was created by experienced teachers.	http://www.azed.gov/assessment/files/2014/05/azella_ft_tech_report_2013_final_0214.pdf	**14 - 15**
			Assessment tasks that are representative of the academic domain can be identified.	Domain Analysis	Assessment tasks were written by experts in the field (WestEd).	http://www.azed.gov/assessment/files/2014/05/azella_ft_tech_report_2013_final_0214.pdf	**14 - 15**
			Assessment tasks that require important skills and are representative of the academic domain can be simulated.	Task Modeling	Authentic comprehension tasks are simulated.	http://www.azed.gov/assessment/files/2014/05/azella_ft_tech_report_2013_final_0214.pdf	**457 - 458**
					PLDs represent the skills needed in the target domain.	http://www.azed.gov/assessment/files/2014/05/az-kpt-standard-setting-report-final_092413.pdf	**3, 7, 12**
					Alignment study demonstrates that the standards are being appropriately sampled.	http://www.azed.gov/assessment/files/2014/04/final-alignment-analysis-azella-report-w-addendum-3-8-14.pdf	**20**

Claim	Warrant	Inference	Assumptions Underlying Warrant	Examples of Backing	Backing for Assumptions	Links to Documentation	Page Number(s)
Observed Score	Observations of performance on test tasks are evaluated to provide observed scores reflective of targeted language abilities.	Evaluation	Rubrics for scoring responses are appropriate for providing evidence of targeted abilities.	Rubric Development	Rubrics are a simple and clear reflection of the skills in the ELPS.	http://www.azed.gov/assessment/files/2014/05/azella_ft_tech_report_2013_final_0214.pdf	**15 - 16**
						http://www.azed.gov/english-language-learners/files/2011/09/stage-i-all.pdf	
			Task administration conditions are appropriate for providing evidence of targeted language abilities.	Prototyping Studies	One-on-one testing provides evidence for appropriate conditions.	http://www.azed.gov/assessment/files/2014/05/azella_ft_tech_report_2013_final_0214.pdf	**14 - 17, 22 - 23**
			The statistical characteristics of items, measures, and test forms are appropriate for criterion referenced decisions.	Item and Test Analysis	Prototyping during item development done in cognitive studies.	http://www.azed.gov/assessment/files/2014/05/azella_ft_tech_report_2013_final_0214.pdf	**14 - 17, 459 - 467**
					Validity studies independently performed nationally by recognized outside vendor (NCEO).	http://www.azed.gov/assessment/files/2014/08/kpt-final-report-8-15-2014.pdf	**4 - 12**
					Statistical characteristics of items were one of the criteria for selection for the test. Statistical characteristics of the full test provide evidence of appropriateness of use.	http://www.azed.gov/assessment/files/2014/06/azella_tech_report_2012-2013_final.pdf	**289, 516**

Claim	Warrant	Inference	Assumptions Underlying Warrant	Examples of Backing	Backing for Assumptions	Links to Documentation	Page Number(s)
Expected Score	Observed scores are estimates of expected scores over the relevant parallel versions of tasks and test forms, and across raters.	Generaliz-ation	A sufficient number of tasks are included on the test to provide stable estimates of test-takers' performances.	Generaliz-ability and Reliability Studies	There is only one version of the test.	http://www.azed.gov/assessment/files/2014/05/azella_ft_tech_report_2013_final_0214.pdf	**1**
					Multiple items included for each task type.	http://www.azed.gov/assessment/files/2014/04/final-alignment-analysis-azella-report-w-addendum-3-8-14.pdf	**20, 29, 32 - 33,**
					Scoring training qualification is required.	http://www.azed.gov/assessment/files/2014/05/azella_ft_tech_report_2013_final_0214.pdf	**36**

Claim	Warrant	Inference	Assumptions Underlying Warrant	Examples of Backing	Backing for Assumptions	Links to Documentation	Page Number(s)
Construct	Expected scores are attributed to a construct of academic and social language proficiency.	Explanation	The linguistic knowledge, processes, and strategies required to successfully complete tasks vary in keeping with theoretical expectations.	Discourse analysis and cognitive processing studies	Use of ELPS and Early Learning Standards establish the construct of academic and social language proficiency.	http://www.azed.gov/assessment/files/2014/05/azella_ft_tech_report_2013_final_0214.pdf	**14 - 16**
			Performance on new test measures relate to performance on other measures of language proficiency as expected theoretically.	Concurrent correlational studies	The construct of English proficiency should be similar between assessments.	http://www.azed.gov/assessment/files/2014/04/concurrent-external-validity-report_final.pdf	**11**
			The internal structure of the test scores is consistent with a theoretical view of language proficiency as a number of highly interrelated components.	Studies of reliability	Internal reliability (Coefficient Alpha)	http://www.azed.gov/assessment/files/2014/06/azella_tech_report_2012-2013_final.pdf	**289**
					preLAS study	http://www.azed.gov/assessment/files/2014/04/concurrent-external-validity-report_final.pdf	**6**
			Test performance varies by student characteristics.	Comparison of studies of group differences	A high percentage of English-Only students should score proficient.	http://www.azed.gov/assessment/files/2014/05/az-kpt-standard-setting-report-final_092413.pdf	**16**

Claim	Warrant	Inference	Assumptions Underlying Warrant	Examples of Backing	Backing for Assumptions	Links to Documentation	Page Number(s)
Target Score	The construct of academic and social language proficiency, as assessed, accounts for the quality of linguistic performance in English-medium Kindergarten classes in Arizona.	Extrapolation	Performance on the test is related to other criteria of language proficiency in the Kindergarten context.	Criterion-related validity studies	Teacher definitions of the academic and social language needed in Arizona Kindergarten classes is included in the NCEO study.	http://www.azed.gov/assessment/files/2014/08/kpt-final-report-8-15-2014.pdf	**37 -38**
					Webb alignment based on standards from Early Childhood and ELPS.	http://www.azed.gov/assessment/files/2014/04/final-alignment-analysis-azella-report-w-addendum-3-8-14.pdf	**iv, 20, 29 32 - 33**

Claim	Warrant	Inference	Assumptions Underlying Warrant	Examples of Backing	Backing for Assumptions	Links to Documentation	Page Number(s)
Decisions	Estimates of the quality of performance in the English-medium Kindergarten classes obtained from the test are useful in making classification decisions regarding whether students are eligible for ELL services in the state of Arizona and about appropriate curriculum for test-takers.	Decision-making	The classification decisions that are made reflect the least acceptable error.	Standard setting studies	Standard Setting	http://www.azed.gov/assessment/files/2014/05/az-kpt-standard-setting-report-final_092413.pdf	15 - 16
					Evidence-based cut-score	http://www.azed.gov/assessment/files/2014/09/kpt-ebssr_report_final.pdf	50, 52
			The meaning of the scores is clearly interpretable by teachers, parents, and school officials.	Score interpretation materials	Performance Level Descriptors	http://www.azed.gov/assessment/files/2014/05/az-kpt-standard-setting-report-final_092413.pdf	3, 7 - 8, 12
					SEM at cut-score	http://www.azed.gov/assessment/files/2014/06/azella_tech_report_2012-2013_final.pdf	512
					Cut-score review	http://www.azed.gov/assessment/files/2014/09/kpt-ebssr_report_final.pdf	8 - 12
					Score interpretation is provided in technical documentation.	http://www.azed.gov/assessment/files/2014/05/az-kpt-standard-setting-report-final_092413.pdf	31
					Score interpretation is provided in the navigating student reports document.	http://www.azed.gov/assessment/files/2015/02/ellreports.pdf	18
					Score interpretation is provided on the student reports.	http://www.azed.gov/assessment/files/2014/05/az-kpt-standard-setting-report-final_092413.pdf	29, 51
			The scores are equitable for children, parents, teachers, and other stakeholders.	Bias studies	Content and Bias review	http://www.azed.gov/assessment/files/2014/05/azella_ft_tech_report_2013_final_0214.pdf	15
					DIF analyses	http://www.azed.gov/assessment/files/2014/06/azella_tech_report_2012-2013_final.pdf	303 - 304

Claim	Warrant	Inference	Assumptions Underlying Warrant	Examples of Backing	Backing for Assumptions	Links to Documentation	Page Number(s)
Consequence	The consequences of using the test and the decisions that are made are beneficial to children, parents, teachers, principals, Arizona's educational system, and other stakeholders.	Policy-based Interpretation	The placement test will have a positive influence in how English is taught both in ELL and mainstream classrooms. It will also positively influence how children perform in Kindergarten and subsequent grades.	Washback studies Examination of classification errors on subsequent performance Examination of subsequent academic performance by IFEP Kindergarten students	Longitudinal studies of students who are assessed using the placement test are being performed both at the state and at the district level.	These studies are in progress.	

Table adapted from the following sources:

Bachman, L., & Palmer, A. (2010). *Language assessment in practice*. New York, NY: Oxford University Press.

Chapelle, C., Enright, M. & Jamieson, J. (Eds.) (2008). *Building a validity argument for the Test of English as a Foreign Language*. Mahwah, NJ: Lawrence Erlbaum Publishers.